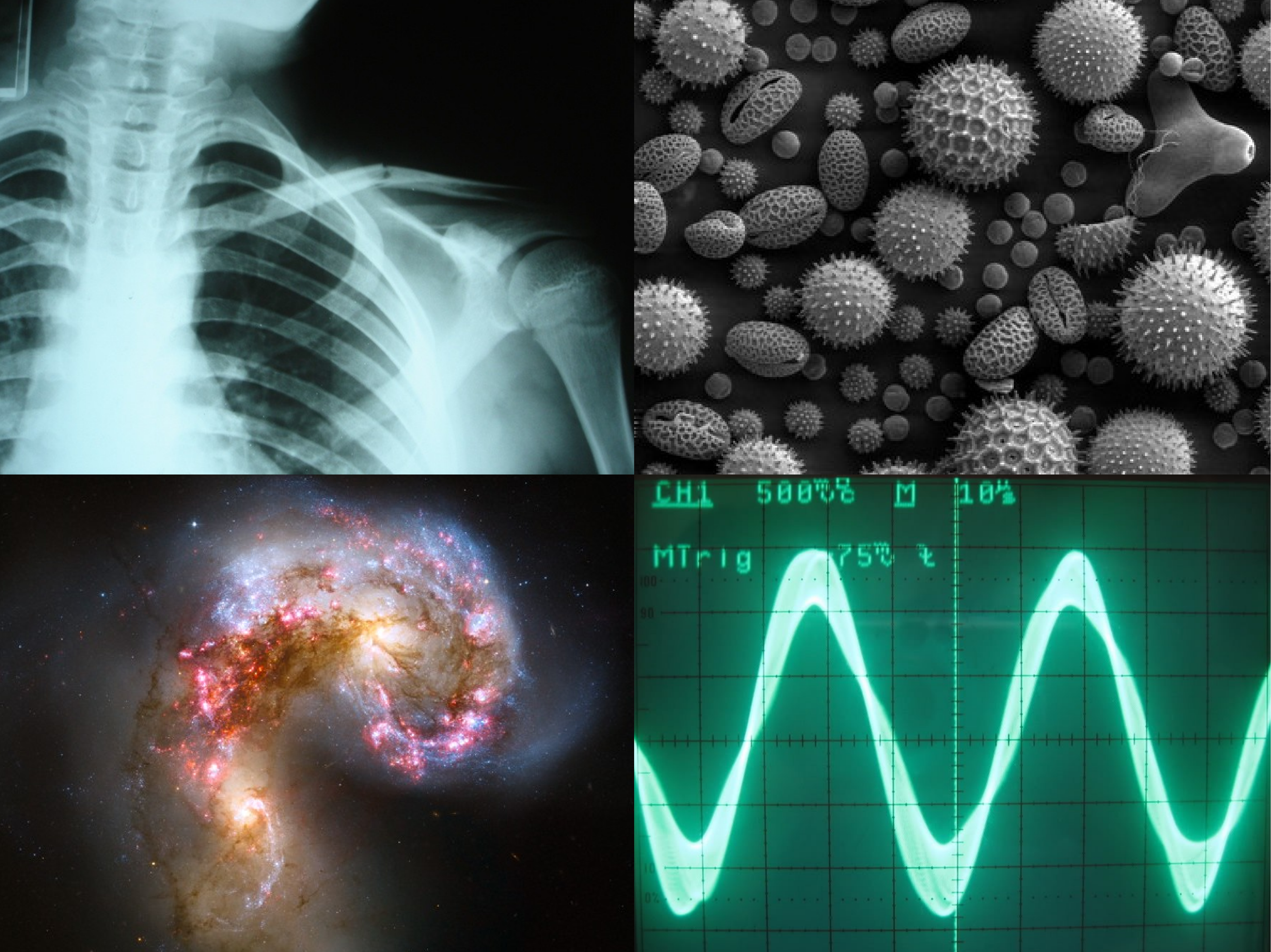
Advanced Trace Analysis

Tracing workshop '11

Francis Giraldeau
francis.giraldeau@polymtl.ca

Under the direction of Michel Dagenais
DORSAL Lab, École Polytechnique de Montréal


CH1
M
MTrig


WWW


Monday 09 May
9 am
10 am
11 am
12 pm
1 pm
2 pm
3 pm
4 pm
5 pm
6 pm
7 pm
10:00am
Datacenter capacity planning
1:30pm
Customer satisfaction damage control

Tracing to the rescue

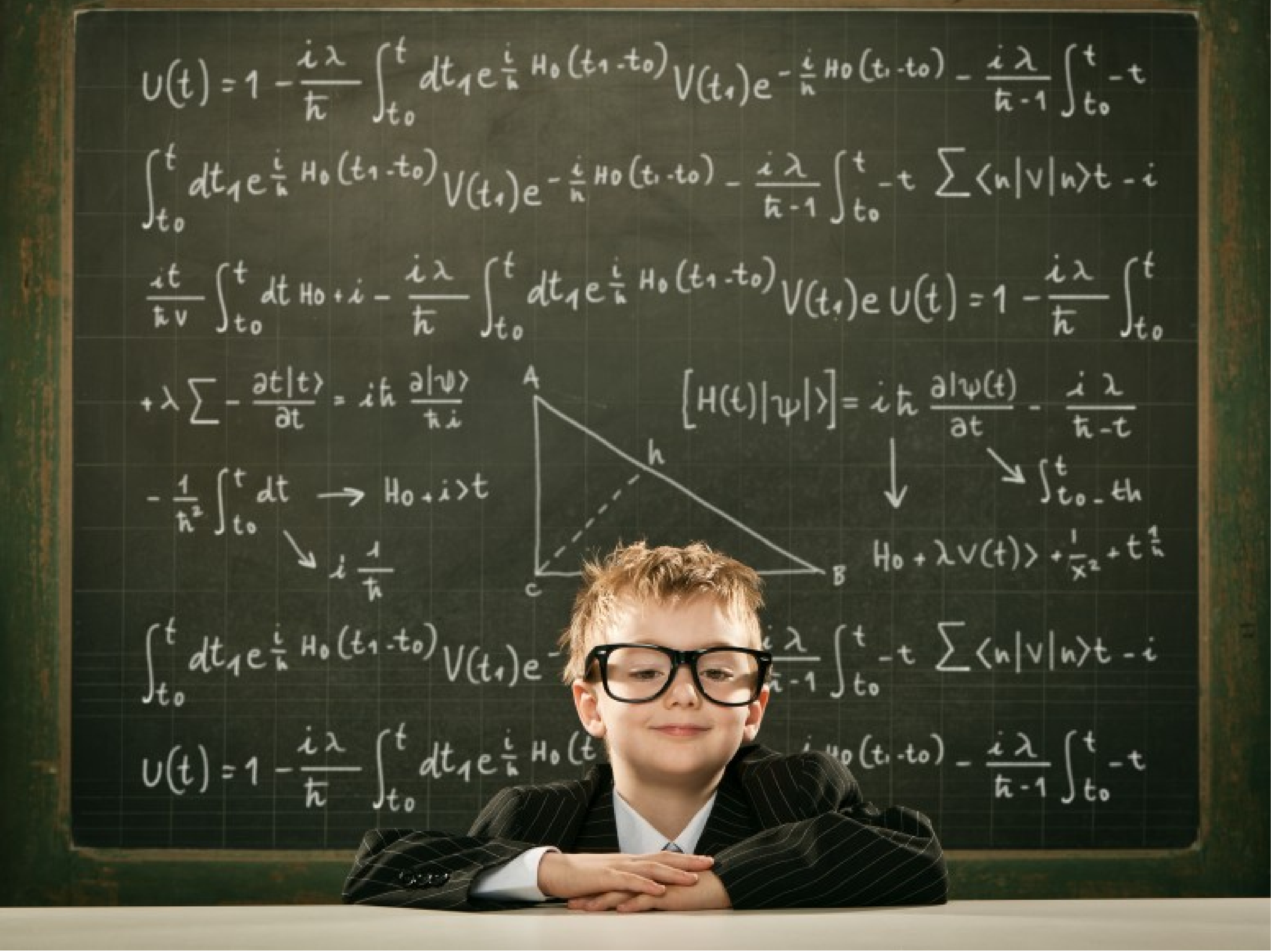
- Highly precise timing information
- Low disturbance
- System wide instrumentation

Challenges

- Requires deep understanding of kernel
- Complex events sequences
- Overwhelming trace size

Current approaches

- Mainstream approaches
  - Events table : display and search events
  - Control flow : state change according to time
  - Histogram : density of events in time
- Dependency analysis module
- Event patterns to abstract low level events
- Event scenarios to search for special conditions


U(t) = 1 - iλ/ħ ∫_{to}^{t} dt1 e^{i/ħ H0 (t1 - to)} V(t1) e^{-i/ħ H0 (t1 - to)} - iλ/(ħ-1) ∫_{to}^{t} -t
∫_{to}^{t} dt1 e^{i/ħ H0 (t1 - to)} V(t1) e^{-i/ħ H0 (t1 - to)} - iλ/(ħ-1) ∫_{to}^{t} -t ∑⟨n|V|n⟩t - i
it/ħV ∫_{to}^{t} dt H0 + i - iλ/ħ ∫_{to}^{t} dt1 e^{i/ħ H0 (t1 - to)} V(t1) e U(t) = 1 - iλ/ħ ∫_{to}^{t}
+ λ ∑ - ∂t|t⟩/∂t = iħ ∂|ψ⟩/ħi
[H(t)|ψ|⟩] = iħ ∂|ψ(t)⟩/∂t - iλ/(ħ-t)
A
h
B
C
- 1/ħ² ∫_{to}^{t} dt → H0 + i⟩t
i 1/ħ
∫_{to - th}^{t}
H0 + λV(t)⟩ + 1/x² + t^{1/h}
∫_{to}^{t} dt1 e^{i/ħ H0 (t1 - to)} V(t1) e
∫_{to}^{t} -t ∑⟨n|V|n⟩t - i
U(t) = 1 - iλ/ħ ∫_{to}^{t} dt1 e^{i/ħ H0 (t
H0 (t1 - to) - iλ/(ħ-1) ∫_{to}^{t} -t

Research objective

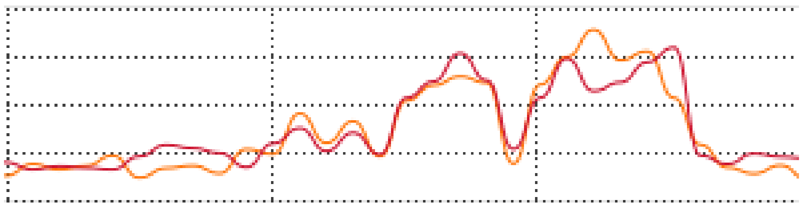
Provide kernel trace analysis algorithms and techniques to allow system administrators and programmers to understand system wide runtime performance behavior of distributed applications.

Tracing for the rest of us

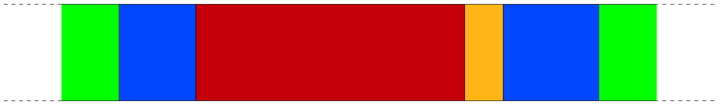
Endless questions...

- What is the critical path of this request?
- Which subsystem is the bottleneck?
- What is the root cause of latency?
- How much resource this task requires?
- What is the relationship between process?

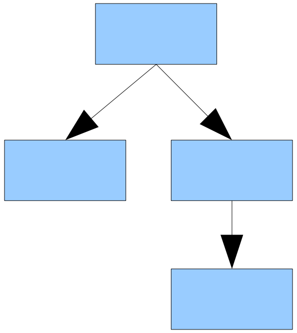
Resource usage recovery



Blocking analysis

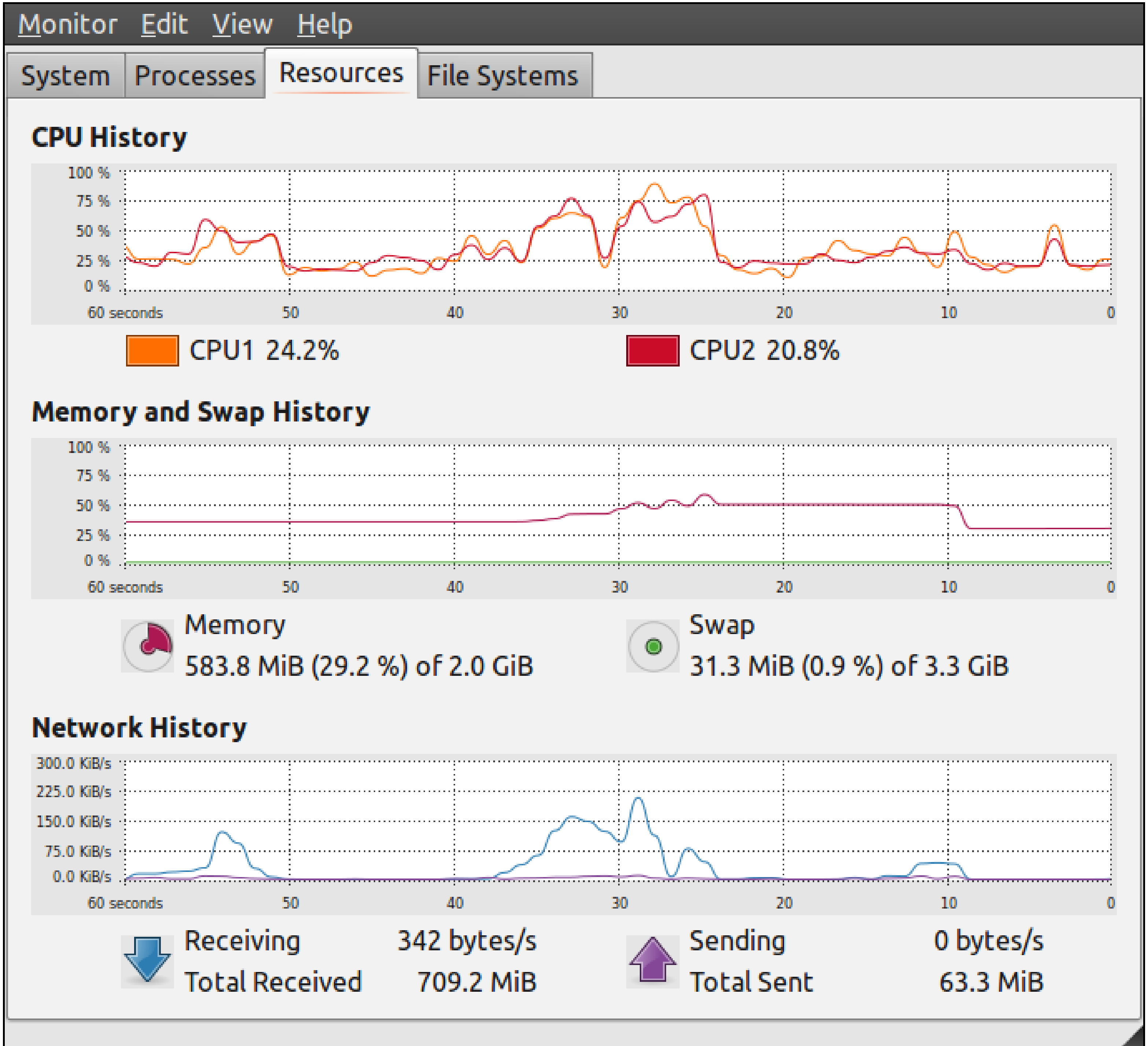


Inter-process relationship recovery

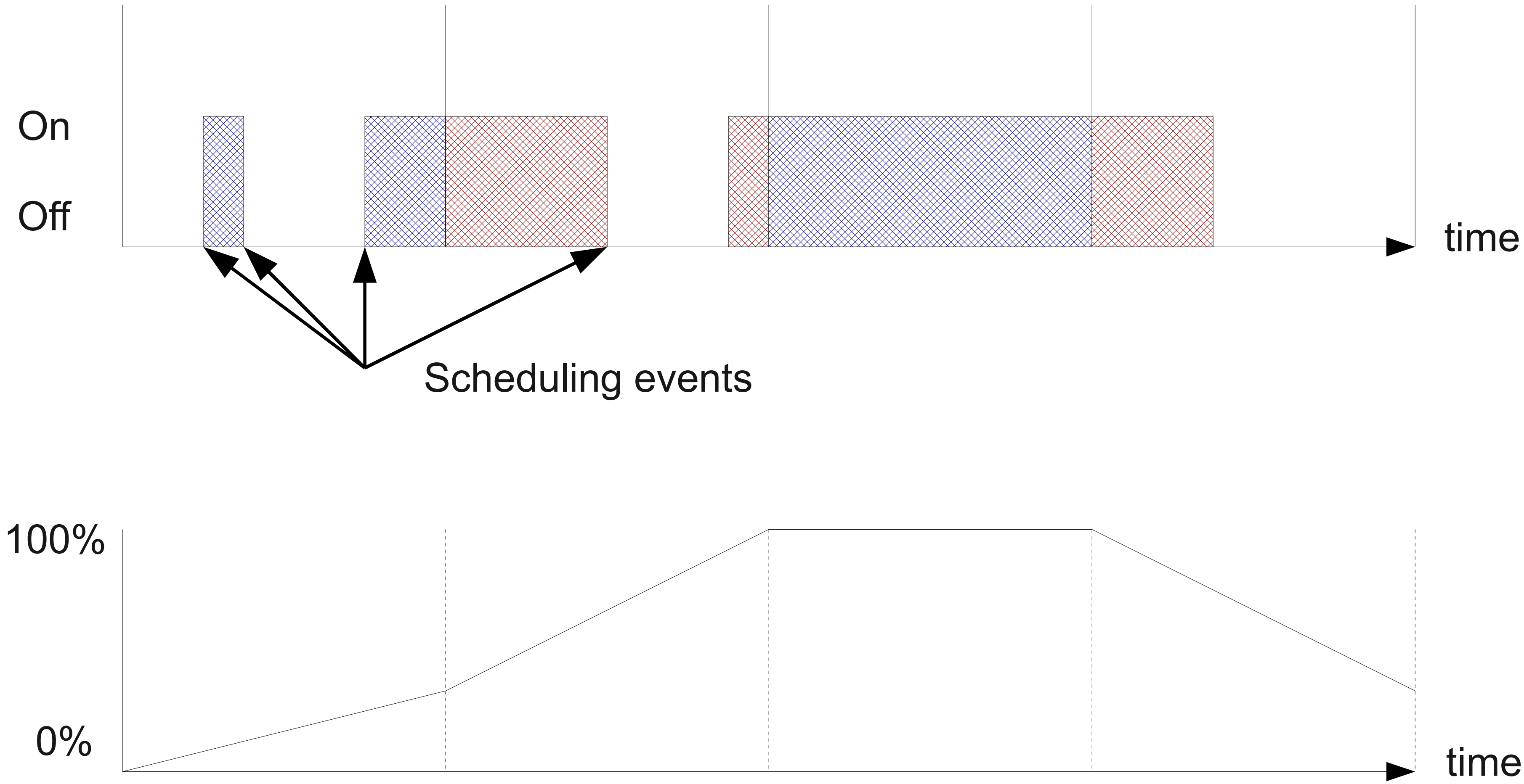


Resource accounting

Resource usage recovery


Monitor Edit View Help
System
Processes
Resources
File Systems
CPU History
100 %
75 %
50 %
25 %
0 %
60 seconds
50
40
30
20
10
0
CPU1 24.2%
CPU2 20.8%
Memory and Swap History
Memory
583.8 MiB (29.2 %) of 2.0 GiB
Swap
31.3 MiB (0.9 %) of 3.3 GiB
Network History
300.0 KiB/s
225.0 KiB/s
150.0 KiB/s
75.0 KiB/s
0.0 KiB/s
Receiving 342 bytes/s
Total Received 709.2 MiB
Sending 0 bytes/s
Total Sent 63.3 MiB

Recovering CPU usage

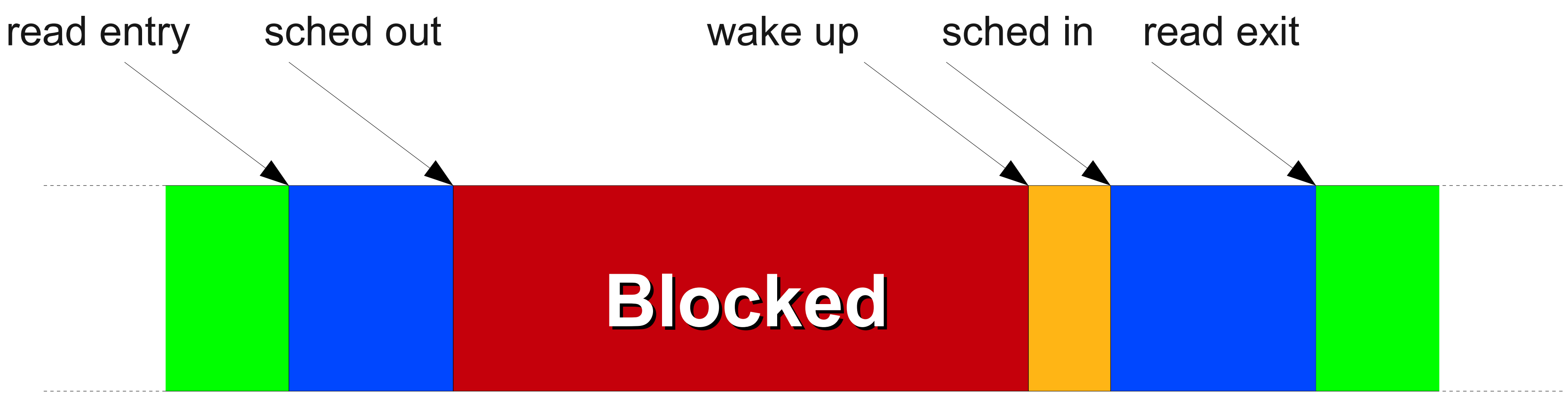


Metrics to recover

- CPU time ✓
- Memory allocations ✓
- Network bandwidth ✓
- Disk I/O ✓

Blocking analysis

Blocking analysis



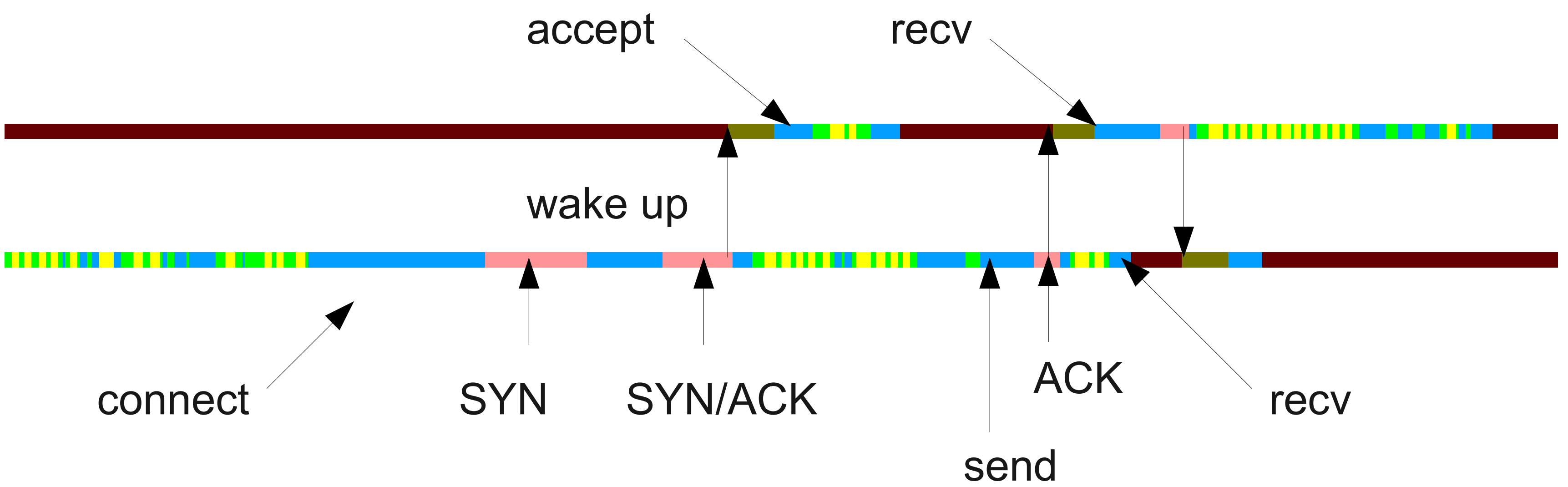
- Always occur in kernel mode
- Presence of wake up event
- What is blocking? How much time is lost?

Inter-process relationship recovery

Inter-process relationship

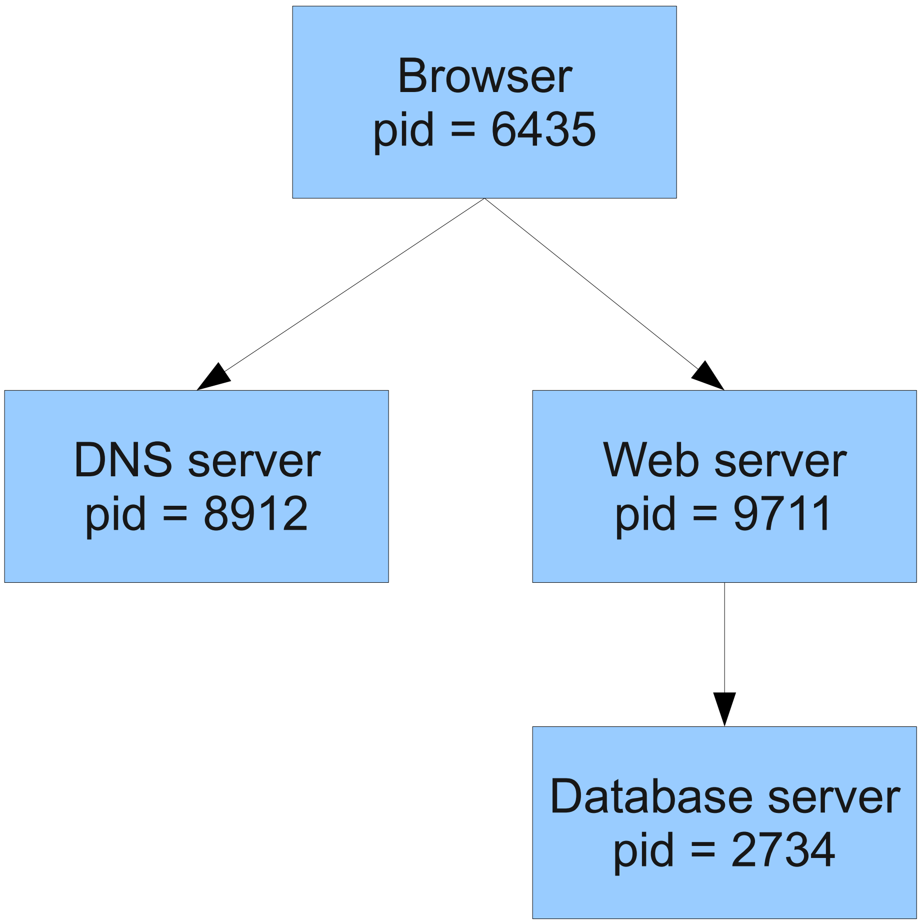
- Unix IPC mechanism
  - Sockets
  - Shared memory
  - Signals
  - Pipe
- File locks
- Futex

TCP sockets recovery



- TCP extended events for IP address and ports
- Socket system calls events (accept, connect)
- SoftIRQ and wake up for incoming packets

Automatic relationship recovery



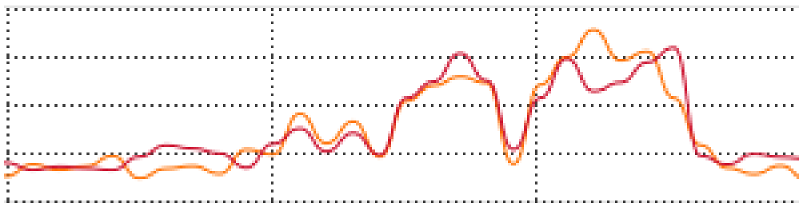
Resource accounting

Resource accounting

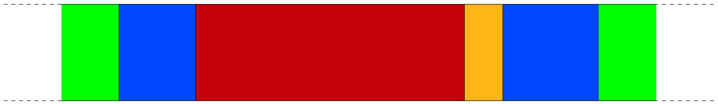
Server is busy on behalf of the client

- Resource usage of subtask should be accounted to the client that does the request
- Inter-process relations + resource usage + blocking analysis

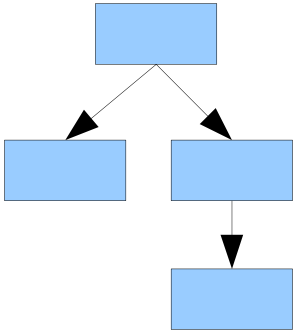
Resource usage recovery



Blocking analysis



Inter-process relationship recovery

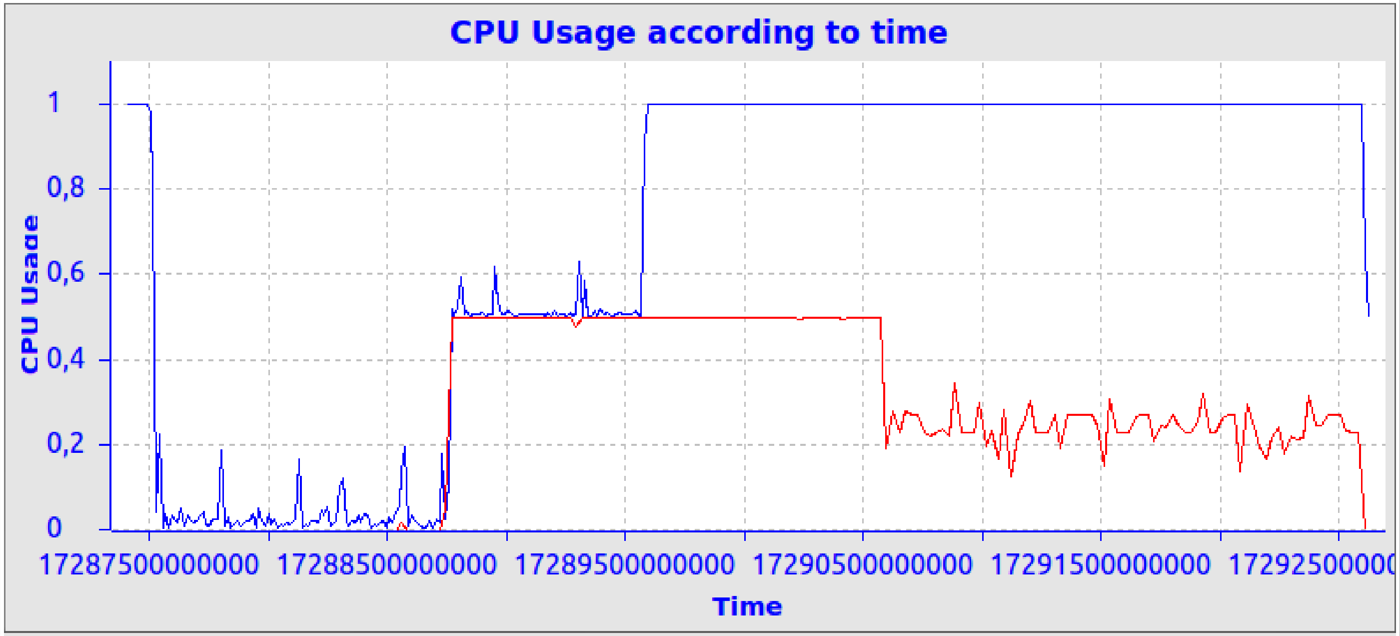


Resource accounting

Demo

File

CPU Usage | Dependency analysis



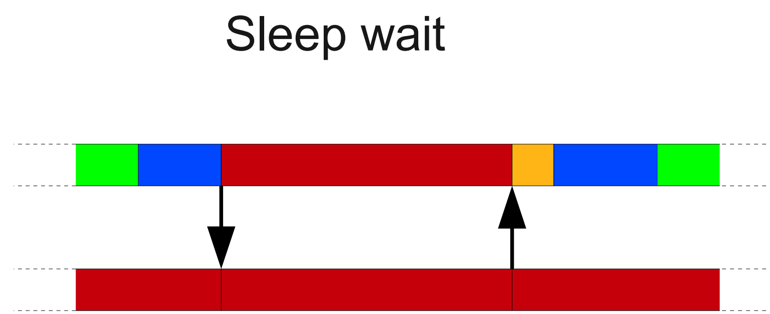
PID	Command	Running
6906	burnP6	2801,931
6908	burnP6	2332,400
6910	burnP6	975,268
1207	Xorg	303,556
2504	evolution	170,106
6899	lttctl	110,802
6901	lttd	86,763
2973	chromium-browse	69,988
6902	lttd	52,380
1971	wnck-applet	47,776
1851	compiz	37,530
2886	chromium-browse	32,796
2806	gnome-terminal	29,525

```
Task pid=7050 cmd=/usr/local/bin/clihog

Systemcall blocking summary              N    Sum (ms)
------------------------------------------------------
sys_read                                 1     100,234


File descriptor blocking summary         N    Sum (ms)
------------------------------------------------------
192.168.0.102:9876                       1     100,234


CPU accounting
PID                                 Self (ms)    Sub (ms)  Total (ms)
------------------------------------------------------------------
7050 clihog                             2,761       0,200       2,961
 \_ 7048 srvhog                         0,200       0,000       0,200
```

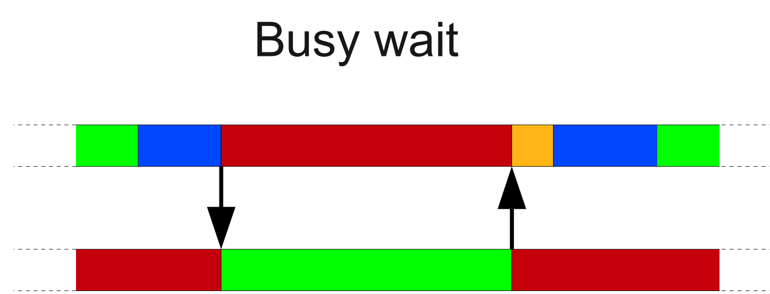



```
Task pid=6986 cmd=/usr/local/bin/clihog

Systemcall blocking summary              N    Sum (ms)
------------------------------------------------------
sys_read                                 2      99,943
ptregs_execve                            1       0,508


File descriptor blocking summary         N    Sum (ms)
------------------------------------------------------
192.168.0.102:9876                       2      99,943


CPU accounting
PID                                 Self (ms)    Sub (ms)  Total (ms)
------------------------------------------------------------------
6986 clihog                             2,504      79,969      82,473
 \_ 6983 srvhog                        79,969       0,000      79,969
```



Future work

- Recover all major metrics
- Complete missing trace events
- State history integration
- Toward cluster-wide live analysis
- Problem with NAT and firewalls
- Combine user space and kernel trace analysis
- LTTng/TMF integration

Finns det någon fråga?

francis.giraldeau@polymtl.ca

http://pages.usherbrooke.ca/fgiraldeau